|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | |  |  | | --- | --- | | For office use only | | | T1 | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | | T2 | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | | T3 | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | | T4 | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | | |  | | --- | | Team Control Number **53478** | |  | | Problem Chosen **C** | | |  |  | | --- | --- | | For office use only | | | F1 | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | | F2 | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | | F3 | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | | F4 | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | |   **2016 MCM/ICM Summary Sheet** (Your team's summary should be included as the first page of your electronic submission.) Type a summary of your results on this page. Do not include the name of your school, advisor, or team members on this page. |

# Introduction

## Problem background

This is a problem about investment.

A charitable organization will donate a certain amount of money to a group of colleges and universities for five years to help improve educational performance of undergraduates in the United States.

We should:

1. Define an appropriate return of investment (ROI) for a charitable organization, which is mainly about it's positive effect on student performance;
2. Rank the candidate schools base on their demonstrated potential for effective use of private funding;
3. Determine an optimal investment strategy that tells which schools to invest, how much to invest on each school, and how long should the duration of investment be, to get the highest ROI;
4. Estimate the ROI of this strategy.

## Previous research

We need to solve the whole problem with a distribution of the investment and a rating system. And we have find some other researches to solve such problem on numerical simulation. One is the A Return on Investment as a Metric for Evaluating Information Systems by the Alexei Botchkarev and Peter Andru. This system provides a comprehensive structure for distribute the investment.

Another model is the tradition investment model; we draw lessons from this model by using qualitative analysis.

We also find some sophisticated machine learning algorithm for our clustering step, the classical model is the k-means++ algorithm by Arthur, D. and Vassilvitskii[2007].

# Assumptions and symbols

## Basic assumptions

We must ignore some factors which have micro effects because we want to build a higher performance model. On the other hand, we need to assume some theorems because our researches are based on machine learning, statistics, data processing theory and investment theory.

1. The data we have is accurate and sufficient
2. The performance of a school and its students is related to the investment to that school.
3. The rate of discount can be ignored so that the same amount of money in the future has the same effect as it is at the period the given data based on.
4. The Attributes that we take into consideration play a vital role in our model.
5. Factors that we ignore do not have an interfere to our model.
6. Deleting some data from the raw data can be allowed because these data have invalid or non-sense data.
7. We can simplify a dataset equivalent to a mathematical expression if we need.
8. The revenue of a school is related to its total net price of all its students.
9. The other income of a school won’t be influenced by the investment to a school.
10. After cluster analysis, the schools in a cluster have some characteristics in common and can predict the performance of each other when investment changes.

## Symbol Description

Table ‑ Symbol Description

|  |  |
| --- | --- |
| Symbol | Description |
|  |  |
|  |  |
| ***Size*** | Number of students in a school |
|  | Total SAT equivalent score of students admitted |
|  | The weight of *(i,d)-th* attribute |
|  | SAT equivalent score of students admitted for *i-th* item |
| ***Comp*** | The completion rate of a school |
| ***Score*** | Average score of students when entering a school |
|  | The *(p-th,j-th)* tuple of the data |
|  | The set of tuples |
|  | The *i-th* centroid of clusters |
|  | A constant for calculating the ***R****i* |
|  | Index for (Attributes, Risks) |
|  | The maximum value among the set |
|  | The minimum value among the set |
| ***Aid****Pell* | Percentage of undergraduates who receive a Pell Grant |
| ***Aid****Loan* | Percent of all undergraduate students receiving a loan |
| ***Locale*** | Locale indicator that indicates the locale condition of a school |
| ***CORR2*** | R-square |
| ***CORR*** | Correlation coefficient |
| ***di*** | The *i-th* item for school score evaluation |
| ***Rg1*** | 150% completion rate for four-year institutions, pooled in two-year rolling averages and suppressed for small n size. For four-year school, students are considered to have graduated "on time" if they graduate within 6 years. |
| ***Rg2*** | 200% completion rate for less-than-four-year institutions, pooled in two-year rolling averages and suppressed for small n size. For two year schools, students are considered to have  graduated "on time" if they graduate within 4 years. |
| ***Rr1*** | First-time, full-time student retention rate at four-year institutions |
| ***Rr2*** | First-time, full-time student retention rate at less-than-four-year institutions |
| ***Rr3*** | First-time, part-time student retention rate at four-year institutions |
| ***Rr4*** | First-time, part-time student retention rate at less-than-four-year institutions |
| ***Ld*** | Predominant degree awarded |
| ***Rp*** | 3-year repayment rate, suppressed for n=30 |
| ***Re*** | Share of students earning over $25,000/year (threshold earnings) 6 years after entry |
| ***Mm*** | Median earnings of students working and not enrolled 10 years after entry |
| ***R*** | Score of a school (the evaluation of a school’s performance) |
| ***NPT*** | Average net price of undergraduates in a school |
| ***Rev*** | Total ***NPT*** of all the undergraduates in a school |
| ***I*** | Total investment of a school |
| ***Conf*** | Confidence of ***I-R*** function |
|  |  |
| ***ΔR*** | Improvement of a school’s score |
| ***ΔI*** | Improvement of a school’s investment |
|  |  |
|  |  |
|  | One of the systematic-risk factor index after normalization |
|  | One of the unsystematic-risk factor index after normalization |
|  | Total utility |
|  | Marginal utility |
|  | Average utility |
|  | Quantity |
|  |  |
|  |  |

# Model design and justification

## Dataset based on

Our model based on the following attributes of the given dataset:

HCM2

CURROPER

SAT\_AVG\_ALL

C150\_4\_POOLED\_SUPP

C200\_L4\_POOLED\_SUPP

PCTPELL

PCTFLOAN

LOCALE

NPT4\_PUB

NPT4\_PRIV

UDGS

DISTANCEONLY

RET\_FTL4

RET\_PT4

RET\_FT4

RET\_PTL4

PREDDEG

RPY\_3YR\_RT\_SUPP

gt\_25k\_p6

md\_earn\_wne\_p10

## Data preprocess

The quality of data affects the quality of the data analysis’s results. In order to help improve the quality of the data and the analysis results, the raw data is preprocessed so as to improve the efficiency and ease of the analysis process. Data preprocessing is one of the most critical steps in our analysis process which deals with the preparation and transformation of the initial dataset. Our data preprocessing methods are divided into following categories:

1. Invalid data cleaning

2. Useless attributes cleaning

3. Some sets of mutual substitution attributes cleaning

4. Data transformation and data reduction

### Invalid data cleaning

First of all, we select data of candidate schools from the whole dataset. Our model won’t care about the schools which are not in the given list.

Furthermore, there are two attributes in the dataset which should be taken into consideration. HCM2 indicates schools that are on Heightened Cash Monitoring 2 by the Department of Education, which means the school has a poor financial situation. CURROPER is a flag for currently operating institution. If a school is closed, we shouldn’t invest to it.

After that, there’re still invalid data.

Data that extracted from large, real-world data and data warehouses can be lack of attribute values or containing only aggregate data. From our dataset, we find a large number of tuples which contains invalid attributes likes ‘NULL’ or ‘PrivacySuppressed’ more than 60% of the total attributes.

Table ‑ Invalid ratio of some attributes

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Attribute | *UGDS* | *PCTFLOAN* | *NPT4\_PUB* | *SAT\_AVG* | *GRAD\_DEBT\_MDN10YR\_SUPP* |
| Invalid Ratio (%) | 9.15 | 9.51 | 81.41 | 81.80 | 18.64 |

Incomplete data can lead to a lots of errors.

We can’t make a credibly assess to a college or university if the attribute is missing. So we make a risk evaluation system, this system will be related in the Model testing and sensitivity analysis chapter. We can get a dataset which have a stable expression in our next steps after deleting some tuples which have an extremely lower security factor value.

### ‘Useless’ attributes cleaning

We consider that our data analysis task will involve data integration, which combines data from multiple sources into a coherent data store, we make sure some attributes of data is useless **for us** because some other attributes have a highly performance when we do the research and can replace of them (useless attributes) completely.

In order to simplify our model, we have to throw away these ‘useless’ attributes. For example, *gt\_25k\_p6* means share of students earning over $25,000/year (threshold earnings) 6 years after entry, we think this conception is one-sided because of the time interval is short and the standard $25,000/year is non-quantitative comparing to the md\_earn\_wne\_p10 which means median earnings of students working and not enrolled 10 years after entry. So we choose the md\_earn\_wne\_p10 to become a quantization standard among the class of future income attributes.

On the other hand, we must consider the generalization of the data, where raw data are replaced by higher level concepts. For example, categorical attributes, like INSTURL, can be generalized to higher concepts, like city or country. Similarly, the numeric attributes’ values (e.g., age or locale) may be mapper to higher level concepts, like Black, White, Hisp and Asian.

Table ‑ Example of attribution merge

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| UGDS\_WHITE | UGDS\_BLACK | UGDS\_HISP | UGDS\_ASIAN | UGDS\_AIAN | UGDS\_NHPI |
| UGDS\_2MOR | | UGDS\_NRA | | UGDS\_UNKN | |
| The attributes about distribution of different specials in American universities or colleges | | | | | |

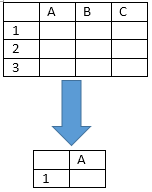


Figure ‑ Data compression from different dimensions

### Some sets of mutual substitution attributes cleaning

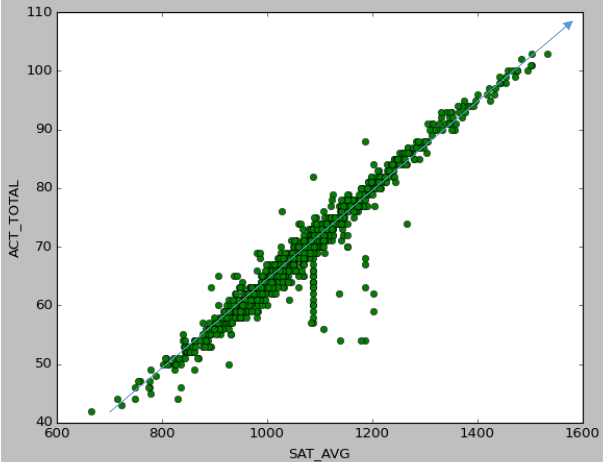


Figure ‑ A linear relationship between SAT scores and ACT scores

As shown in Table 3‑4, it shows the relationship between SAT scores and ACT scores we can conclude that the ACT scores and SAT scores which express the average of the different colleges or universities have a linear relationship. So we can use the ACT\_TOTAL (Average SAT equivalent score of students admitted) or SAT\_TOTAL(Total SAT score of students admitted) to represent the undergraduates’ entrance examination results.

Like this, we can analysis that some other pairs of attributes have such likely features, such as gt\_25k\_p6 and md\_earn\_wne\_p10. We can use this property to make sure the critical factors to our analysis model.

### Data transformation and reduction

In data transformation and reduction, the data are transformed into appropriate forms for the cluster analysis, information mining or attribute correlation. In our related works, we mainly do three ways of transformation:

1. Transform the attribute’s value to a **normalization** index

We utilize a simple but efficient function to make the attribute’s value clarity. The attribute data are scaled so as to fall within a small specified range 0 to 1(float number).

Obviously, the maximum one’s index is 1, and the minimum one’s index is 0. We realize that this way can give a clearly description for our programs.

2. **Aggregation** the data

We use the summary or aggregation operation are applied to the data. For a instance, the items such as *ACTENMID* (Midpoint of the ACT English score), *ACTMTMID* (Midpoint of the ACT math score) and *ACTWRMID* (Midpoint of the ACT writing score) can summarize a newly *ACT\_TOTAL* to weigh the concerns of the ACT scores.

Like this, we can reduce the data into a generality index to measure the different influence factor and decrease the amount of the attributes.

3. Use transformation to deal with the missing data

The data in the given dataset is not complete and a lot of schools have a NULL for their SAT and ACT scores. For these schools, we have to find another indicator to evaluate the quality of students: ***Score***.

After analyzing the data, we found completion rate of a school is related to the quality of students.

C150\_4\_POOLED\_SUPP and C200\_L4\_POOLED\_SUPP can be merged as **Comp**. That is, for four-year institutions, **Comp** = C150\_4\_POOLED\_SUPP. And for less-than-four-year institutions. **Comp** = C200\_L4\_POOLED\_SUPP.

We draw its trend line and calculated its equation:

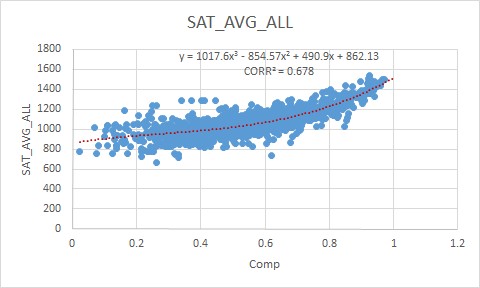


Figure ‑ Scatter of *Comp* and *SAT \_AVG\_ALL* (all candidate schools)

From Figure 3‑3 we can see that the schools with ***Comp*** less than 0.3 has less data than the rest. More exactly, for 849 schools of ***Comp*** less than 0.3, only 124 of them has data for SAT\_AVG\_ALL, and for 1964 schools of ***Comp*** larger than 0.3, 1289 of them has data for SAT\_AVG\_ALL.

So we ignore these data and have another scatter.

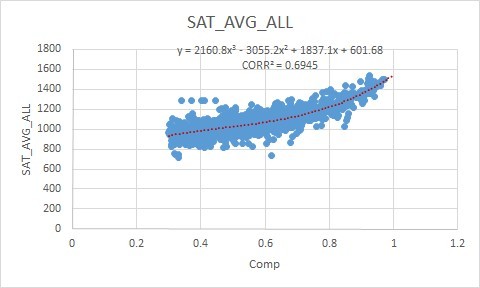


Figure ‑ Scatter of *Comp* and *SAT \_AVG\_ALL* (schools with *Comp* larger than 0.3)

The *CORR2* of the second curve is larger than the first one, indicating that the relationship are more evident.

Then, we use ***Comp*** to calculate ***Score***. If a school has the data of SAT score,

***Score*** = ***SAT \_AVG\_ALL***

If it’s NULL,

## Cluster analysis

Cluster analysis contains a broad suite of designed to find groups of similar attributes within a dataset. Partitioning methods divide the data set into a number of groups predesigned by the user. In our model, we must do a coarse-grained classification to simplify our system. After the process of cluster analysis, we will do qualitative analysis to our clusters and research the distance between the various clusters.

We choose the machine learning algorithm to complete our research because machine learning has three advantages: effectiveness, online learning ability and efficiency.

Between the all kinds of outstanding algorithms in machine learning, we find an almost excellent algorithm that which is called k-means clustering algorithm. k-means clustering that is a method of [vector quantization](https://en.wikipedia.org/wiki/Vector_quantization), aims to [partition](https://en.wikipedia.org/wiki/Partition_of_a_set) n observations into k clusters in which each observation belongs to the cluster with the nearest [mean](https://en.wikipedia.org/wiki/Mean), serving as a [prototype](https://en.wikipedia.org/wiki/Prototype) of the cluster.

We can describe our algorithm in a concise way containing three steps:

1. Give an initial set of **k** means , the algorithm proceeds to next two steps.

2. Assign each observation to cluster whose mean yields the least within cluster sum of squares (WCSS). This is the nearest mean after calculating the squared **Euclidean distance** (Assign different weight to different dimensions)

3. Calculate the new means to be the centroids of the observations in the new clusters. Then do the second step until the clusters tend to be stable.

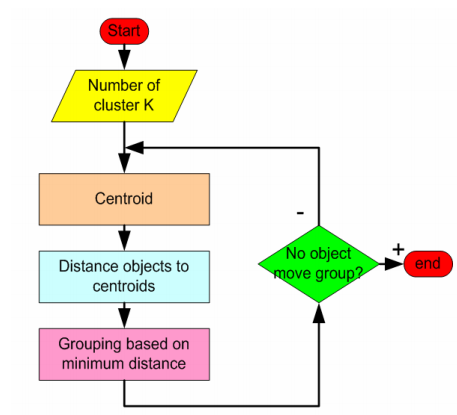


Figure ‑ Flow chart of k-means algorithm

In order to take an optimal **K-value**, we use an improvement algorithm of k-means algorithm. This kind of improvement algorithm called **k-means++** chooses initial centers in a way that gives a provable upper bound on the cluster number. After the running of our program, we find an upper bound between 4 and 6, so we assign an integer 4 to K.

We have run the program we have designed above, and gained two useful datasets.

1. the attribute values of centroid of clusters.

2. the K sets of different university or colleges.

We running this program several times because we randomly initialize the clusters’ centroid when the first step of our program. And then, we received a set of average value after handling the several similar data. We make sure the different automatic initialization can gain an identical result. From Table 3‑3, we can infer the number of universities or colleges in clusters have a similar quantity.

Table ‑ Number of schools in clusters

|  |  |  |  |
| --- | --- | --- | --- |
| Time  cluster | 1 | 2 | 3 |
| 1 | 799 | 838 | 969 |
| 2 | 1198 | 1157 | 1034 |
| 3 | 185 | 266 | 351 |
| 4 | 634 | 555 | 462 |

## The explanation and analysis of the clusters

The attributes for our cluster analysis are:

***Score:*** Average score of students when entering a school.

***Aid****Pell* ***:*** Percentage of undergraduates who receive a Pell Grant, named as *PCTPELL* in the given database

***Aid****Loan* ***:*** Percent of all undergraduate students receiving a loan, named as *PCTFLOAN* in the given database

***Locale:*** Locale indicator that indicates the locale condition of a school. As *LOCALE* in given database is an integer and the larger a locale is, the smaller the integer is. So we did some transformation. That is:

After cluster analysis we get 4 clusters, representing 4 kinds of schools.

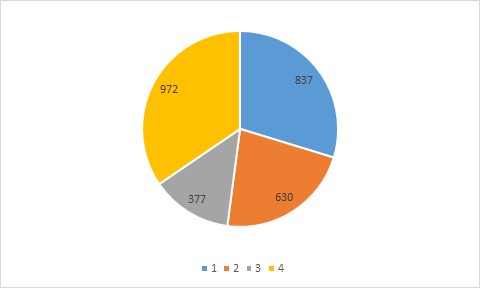


Figure ‑ Number of schools in each cluster

Table 3‑4 shows the centers of these clusters.

Table ‑ Centers of the four clusters

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| cluster | ***Score*** | ***Aid****Pell* | ***Aid****Loan* | ***Locale*** |
| 1 | 1066.30 | 0.59 | 0.39 | 6.13 |
| 2 | 841.37 | 0.34 | 0.48 | 5.97 |
| 3 | 1255.41 | 0.47 | 0.26 | 6.57 |
| 4 | 958.12 | 0.43 | 0.44 | 4.69 |

As we can see, the first and the third cluster has a higher quality of students, and is locates in a larger place. While the difference between them is that the students in the first cluster applied for more financial aids. The schools in the second and forth clusters may locates in small towns or so. They may need more help to improve the quality of education.

Schools in the same cluster may have more in common and can indicate the future development of each other.

## Marginal efficiency of investment

The marginal efficiency of investment is that [rate of discount](https://en.wikipedia.org/wiki/Discounting) which would equate the price of a [fixed](https://en.wikipedia.org/wiki/Fixed_asset) investment [asset](https://en.wikipedia.org/wiki/Asset) with its [present discounted value of expected income](https://en.wikipedia.org/wiki/Discounted_cash_flow). We must consider the performance of our investment, so we need to introduce the marginal efficiency to our model.

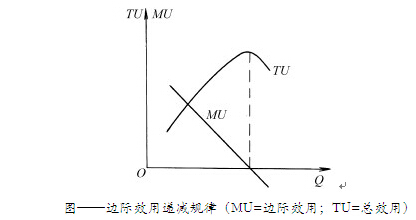


Figure ‑ Marginal efficiency model

***MU***= marginal utility, ***TU***=total utility, ***AU***=average utility.

As a simple example, we will face a law of diminishing marginal utility if we invest a small amount. On the other hand, we must think over the law of increasing marginal utility within an excessive investment. We consider the investment needs to be an appropriate amount. This is because the ***MU*** will returns to negative with increasing of ***Q*** (quantity). These variables can be expressed as follows:

So we must give a reasonable distribution for our investment, we cannot make an unbalance distribution occurs to ***MU*** low to a certain extent. Or we will have a failed scheme for our distribution.

## Evaluate the schools

### R-the quantized metric of schools

Explanation

The ROI is supposed to be the return of investment. In this case it is served as a quantized metric to predict how the quality of a school is likely to change by given a certain amount of investment I. Hence we need a new quantized metric to indicate the quality of school in order to estimate the change of school quality.

we use the following model in calculation of ***R***.

### The Weighted average model for evaluation

Inspired by the methods school ranking organizations like QS World University Rankings

[http://www.topuniversities.com/university-rankings-articles/world-university-rankings/qs-world-university-rankings-methodology]

and ARWU-Methodology

[http://www.topuniversities.com/university-rankings-articles/world-university-rankings/qs-world-university-rankings-methodology]

#### The evaluation model

We adopted a multi-aspect model to evaluate the quality of university. We evaluate school from 3 aspects below, each aspect is evaluated by scoring the data belongs to the aspect.

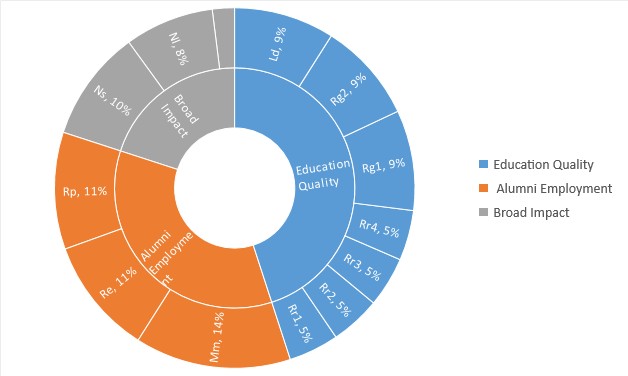


Figure ‑ The percentage of the aspects to evaluate a school

1. Educational quality

This aspect indicates the educational performance of a school.

We evaluate this aspect by calculating the weighted average of the score from the following specs (remember those involved in calculation of ***R*** are scores evaluated with the method in 3.6.2.2 instead of raw data)

***Rg1*** and ***Rg2***, with a weight 0.09, named as *C150\_4\_POOLED\_SUPP* and *C200\_L4\_POOLED\_SUPP* in database respectively.

These specs are about how likely the students in the school can complete their education.

***Rr1***, ***Rr2***, ***Rr3***, ***Rr4***, each with a weight 0.045, named as RET*\_FTL4*, *RET\_PT4*, *RET\_FT4*, *RET\_PTL4* in database respectively.

this spec can be affect by non-educational issues (like illness, poverty, etc.), so the weight is low.

***Ld***, with a weight 0.09, named as *PREDDEG* in database.

this spec indicates the predominant degree awarded research capability of a school, so it is Weighed highly.

1. Alumni Employment

This aspect indicates how likely the student graduated from the university may have a bright future. Though incoming doesn't necessarily mean a bright future, but it does indicate something.

***Rp***, with a weight 0.105, named as *RPY\_3YR\_RT\_SUPP* in database.

How likely the student can pay back the loan in 3 years. This can somehow indicate the employment status of student right after their graduation.

***Re***, with a weight 0.105, named as *gt\_25k\_p6* in database.

The percentages of Alumni who earns more than 25k per year, which are potentially in elite class.

***Mm***, with a weight 0.14, named as *md\_earn\_wne\_p10* in database.

The median of 1ncome from the students graduated for 10 years, the general incoming level measurement.

1. Social Impact

This indicator shows an estimation of the extent the school affect the society

***Ns*** with a weight 0.1, named *UDG*S in database.

This tells the student number of the school. It measures the number of students the school sent to society and can somehow indicate the fame of school since famous schools are always tend to be large and attractive to students.

***Id*** with a weight 0.02, named as *DISTANCEONLY* in database.

indicates if the school only give lecture via internet, an internet only university are not likely to make broad impacts.

***Nl*** with a weight 0.08, named as *LOCALE* in database.

The scale of location where the school is located. General school in rural and remote districts are less likely like to affect the society in large extent.

#### The approximated model of data item score evaluation

For data items ***di*** uniformly distributed, we use the value of the data item directly and then uniform it to [0,1]

For non-negative data items ***di*** distributed as a J-curve, we perform logarithm to it to convert it to a uniformly distribution.

#### Normalization

We adopted the most common way to normalize:

#### Null values and **W**

The given data of schools are sometimes incomplete for reasons like privacy or just missing, the database holds null values. To measure a university as accurate as possible, we employed the sum of weight ***W*** as a measurement of data integration and the scalar of score for a university. We use as the final score of a school. Incomplete data means uncertainty which always brings extra risk too.  
In short, the performance of a schools:

## Return of investment

### Quantify the investment

The current investment of each school cannot be find in the given dataset, and another indicator should be use to quality the investment.

NPT4\_PUB and NPT4\_PRIV in the given database means average net price for Title IV institutions. It’s the average annual total cost of attendance, including tuition and fees, books and supplies, and living expenses, minus the average grant/scholarship aid. It is calculated for all full-time, first-time, degree/certificate-seeking undergraduates who receive Title IV aid. And it can be a measurement of a school’s financial revenue. We assume that the revenue of a school is related to its total net price of all its students.

***NPT*** is theaverage net price of undergraduates in a school. If it’s a public institution, ***NPT*** = NPT4\_PUB; if it’s a private institution, ***NPT*** = NPT4\_PRIV;

Different school has different **Size** (UGDS in the database) so the total revenue

Here we also use ***Rev*** to quantify the further investment to schools, which is ***I***.

### Relationship between investment and return

Shows the result of cluster analysis of these schools, and the schools in every cluster has their own characteristic. The following figures show the scatters of schools in each cluster to see the relationship between ***I*** and ***R***. We assume that after cluster analysis, the schools in a cluster have some characteristics in common and can predict the performance of each other when investment changes.

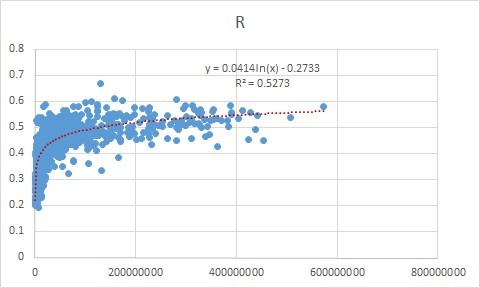


Figure ‑ Relationship between *I* and *R* of schools in cluster 1

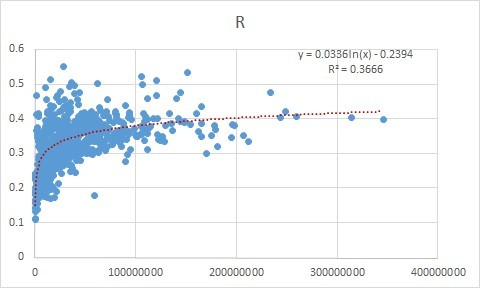


Figure ‑ Relationship between *I* and *R* of schools in cluster 2

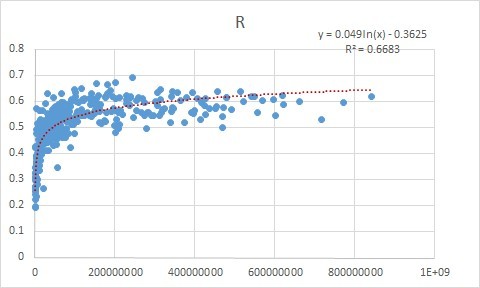


Figure ‑ Relationship between *I* and *R* of schools in cluster 3

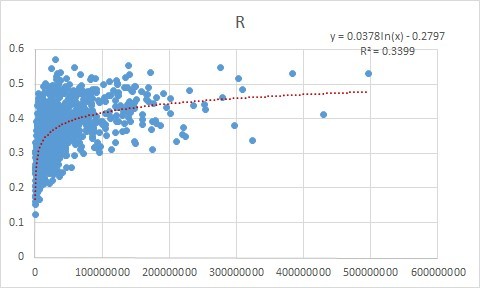


Figure ‑ Relationship between *I* and *R* of schools in cluster 4

As is shown in the above figures, the trend line of ***I*** and ***R*** fits in a logarithmic curve and different clusters have different equations.

What’s more, the ***CORR*** of the trend line is also different. Cluster 2 has the highest ***CORR*** of 0.8187, while cluster 3 has the lowest of 0.6914.

Thus, confidence of school ***j*** in cluster ***i*** can be calculated as following:

## Develop the investment strategy

# Model testing and sensitivity analysis

## Model Test

## Risk and security analysis

Large and complex investment projects have always needed a substantial management structure to insure that our different parts in an organized fashion to achieve the tasks at hand. We think that different university or colleges like different stocks. And anyone of the ‘stocks’ exists **systematic risks** and **unsystematic risks**. So we design a model to calculate the risk-factor index to evaluate the risks in a comprehensive method.

We can consider lacking of vital data is a kind of systematic risk, so we evaluate various attributions to conclude an experimental formula and assign weight to the factors. Then we analysis two sides of risk and give an appropriate value 0.4 for coefficient of ***d***.

From the following figures, we can gain a growth curve that describes the increasing of universities or colleges’ total quantity by the increasing of risk-factor index. Our figure is similar with the standard risk growth curve in most of the models.

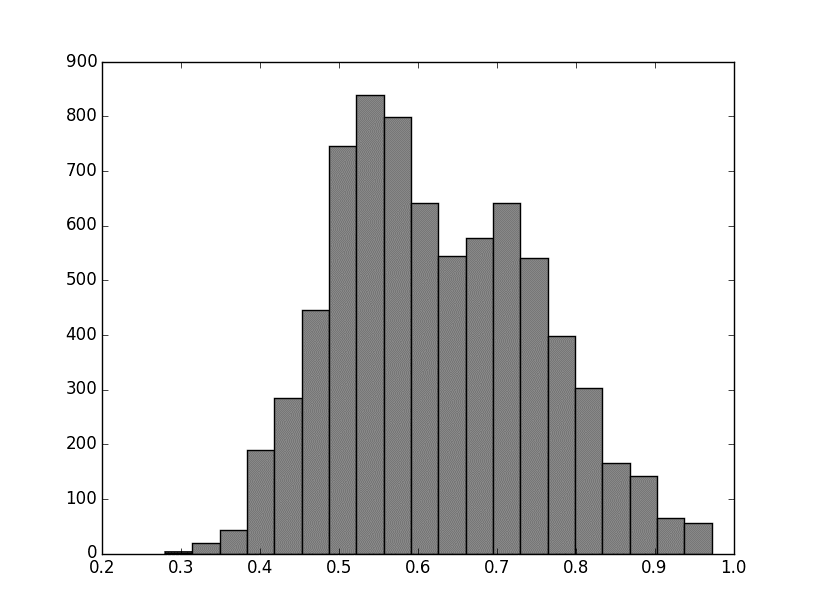


Figure ‑ The risk distribution of index numerical difference between 0.2 to 1.0

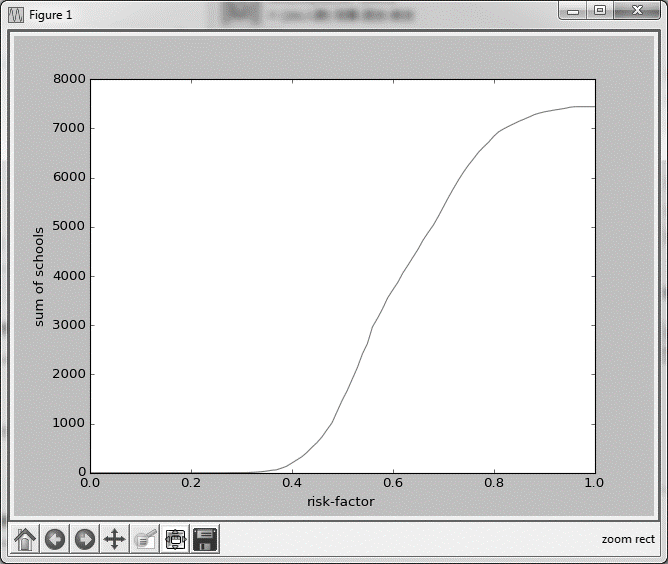


Figure ‑2 The growth curve of sum of the number of schools and risk factor index

Table ‑ weight of attributes

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Attribute | HCM2 | UGDS | SAT\_AVG\_ALL | md\_earn\_wne\_p10 | GRAD\_DEBT\_MDN\_SUPP |
| Weight | 7 | 3 | 3 | 2 | 2 |
| Attribute | NPT4 | PCT | UG25abv | POOLED\_SUPP | RPY\_3YR\_RT\_SUPP |
| Weight | 1 | 1 | 2 | 4 | 4 |

With the systematic risks, we give a set of weights to the major ten attributes which play a vital role in our model. So we can estimate an appropriate value of the systematic risks.

With the unsystematic risks, we choose *RPY\_3YR\_RT\_SUPP* (3-year repayment rate, suppressed for n=30) and *POOLED\_SUPP* (a new attribute measures the graduation rate by merging *C150\_4\_POOLED\_SUPP* and *C200\_L4\_POOLED\_SUPP* in the previous chapter) to become our measure standard, and we adopt an equal weight method to estimate the unsystematic risks.

# Conclusions

The list of schools and the amount to grant in 5 years

The plan is about to grant to the following schools continuously for five years, each year the same amount to achieve ROI of 0.311214 per year.

Table ‑ School list and amount

| School unit ids | School name | Amount |
| --- | --- | --- |
| 101897 | Northeast Alabama Community College | 6030000 |
| 140678 | North Georgia Technical College | 5964000 |
| 203331 | Eastern Gateway Community College | 4872000 |
| 198507 | Elizabeth City State University | 4620000 |
| 366465 | Ogeechee Technical College | 4576000 |
| 196699 | Ulster County Community College | 4400000 |
| 240505 | Casper College | 4224000 |
| 102076 | Snead State Community College | 4120000 |
| 171395 | North Central Michigan College | 4080000 |
| 176008 | Mississippi Delta Community College | 4032000 |
| 141255 | Wiregrass Georgia Technical College | 3652000 |
| 234377 | Wytheville Community College | 3618000 |
| 171225 | Monroe County Community College | 3612000 |
| 217712 | Technical College of the Lowcountry | 3528000 |
| 206923 | Carl Albert State College | 3150000 |
| 117195 | Lake Tahoe Community College | 3050000 |
| 176239 | Pearl River Community College | 2954000 |
| 236072 | Seattle Community College-North Campus | 2698000 |
| 175573 | Copiah-Lincoln Community College | 2652000 |
| 121363 | Porterville College | 2640000 |
| 104577 | Eastern Arizona College | 2568000 |
| 135160 | Florida Gateway College | 2450000 |
| 199926 | Wilkes Community College | 1960000 |
| 138682 | Albany Technical College | 1808000 |
| 129756 | Middlesex Community College | 1800000 |
| 207306 | Northwestern Oklahoma State University | 1682000 |
| 137315 | South Florida State College | 1512000 |
| 187958 | University of New Mexico-Gallup Campus | 1406000 |
| 141811 | Leeward Community College | 1352000 |
| 198321 | Cleveland Community College | 1224000 |
| 175643 | East Central Community College | 1080000 |
| 101286 | George C Wallace State Community College-Dothan | 1008000 |
| 139010 | Bainbridge State College | 972000 |
| 121707 | College of the Redwoods | 406000 |
| 233019 | Patrick Henry Community College | 300000 |

# Strengths and weaknesses

## Strengths

1. **Higher performance**: we writing some scripts by using Python C# and Matlab to make the whole processing more automated. After that, we use a version control system to solve the conflicts.
2. **Simplicity**: the programs of the model consists of the basic mathematical methods. And we simplify our model as far as possible without lacking of accuracy.
3. **Accuracy**: we make use of the fine grained model in most of our research. And we make great efforts to control the precision of the data.
4. **Adaptability and Practicability**: the model we build has good portability; it is suitable for other likely models.

## Weaknesses

We did not consider some other factors which can also have an influence to us. So we choose a research way to ignore some attributes of the dataset. In some sense, our ‘accuracy’ means have a correct result under the existing model.

# A letter to the CFO of Goodgrant Foundation

Dear *Mr. Alpha Chiang*:

It is my pleasure to write to you.

The strategy we use is to maximize ROI while maintaining a low investment risk.

We model the investment by two aspects: Return and Risk.

We firstly preprocessed our data with various categories of methods and classified the schools to four clusters according to their potential consistency by applying a k-means clustering algorithm to the schools. Then we constructed a model to quantify return and by using the total expense as an estimation of the total revenue, we managed to construct the model between return and investment by fitting the return-revenue curve for each class. We got a relatively reliable logarithm model set for each class. After that we filtered out schools obviously too risky for investment including but not limited to those not being in operation, with HCM2 flag or having too few students. Finally, the remaining 1601 candidates are put into a combination optimization model, which takes both ROI and risk brought by data incompetence into consideration, an optimal solution with 25 candidates with different amount of grant is found.

The following table contains the schools and the amounts respectively.

| school unit ids | school name | amount |
| --- | --- | --- |
| 101897 | Northeast Alabama Community College | 6030000 |
| 140678 | North Georgia Technical College | 5964000 |
| 203331 | Eastern Gateway Community College | 4872000 |
| 198507 | Elizabeth City State University | 4620000 |
| 366465 | Ogeechee Technical College | 4576000 |
| 196699 | Ulster County Community College | 4400000 |
| 240505 | Casper College | 4224000 |
| 102076 | Snead State Community College | 4120000 |
| 171395 | North Central Michigan College | 4080000 |
| 176008 | Mississippi Delta Community College | 4032000 |
| 141255 | Wiregrass Georgia Technical College | 3652000 |
| 234377 | Wytheville Community College | 3618000 |
| 171225 | Monroe County Community College | 3612000 |
| 217712 | Technical College of the Lowcountry | 3528000 |
| 206923 | Carl Albert State College | 3150000 |
| 117195 | Lake Tahoe Community College | 3050000 |
| 176239 | Pearl River Community College | 2954000 |
| 236072 | Seattle Community College-North Campus | 2698000 |
| 175573 | Copiah-Lincoln Community College | 2652000 |
| 121363 | Porterville College | 2640000 |
| 104577 | Eastern Arizona College | 2568000 |
| 135160 | Florida Gateway College | 2450000 |
| 199926 | Wilkes Community College | 1960000 |
| 138682 | Albany Technical College | 1808000 |
| 129756 | Middlesex Community College | 1800000 |
| 207306 | Northwestern Oklahoma State University | 1682000 |
| 137315 | South Florida State College | 1512000 |
| 187958 | University of New Mexico-Gallup Campus | 1406000 |
| 141811 | Leeward Community College | 1352000 |
| 198321 | Cleveland Community College | 1224000 |
| 175643 | East Central Community College | 1080000 |
| 101286 | George C Wallace State Community College-Dothan | 1008000 |
| 139010 | Bainbridge State College | 972000 |
| 121707 | College of the Redwoods | 406000 |
| 233019 | Patrick Henry Community College | 300000 |

The ROI we proposed is defined as the incremental of the quantified return R caused by the investment divides the incremental of the revenue caused by Investment.

Look forward to your feedbacks and suggestions soon.

*Sincerely,*

*A MCM Team*

# Appendix

# Reference